# A ROBUST SCRIPT IDENTIFICATION SYSTEM FOR HISTORICAL INDIAN DOCUMENT IMAGES

***S. Kavitha[1], P. Shivakumara[2], G. Hemantha Kumar[3] and C. L. Tan[4]***

[1, 3] Department of Studies in Computer Science, University of Mysore-Karnataka, India

[2]Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.

[4] School of Computing, National University of Singapore, Singapore.

Email: [1]kavitha_sanjay_as@yahoo.co.in, [2]shiva@um.edu.my, [3]ghk.2007@yahoo.com, [4]tancl@comp.nus.edu.sg

## *Abstract*

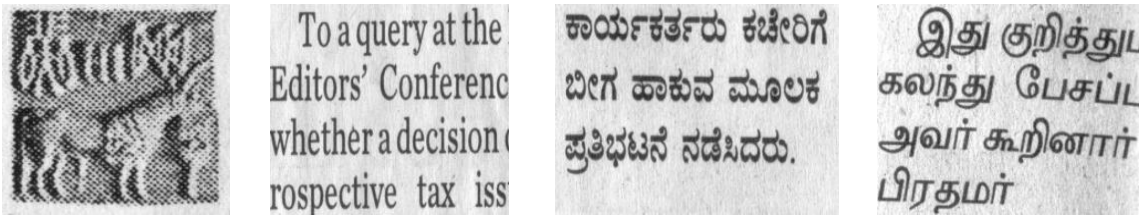
*Automatic script identification in archives of documents is essential for searching a specific document in order to choose an appropriate Optical Character Recognizer (OCR) for recognition. Besides, identification of one of the oldest historical documents such as Indus scripts is challenging and interesting because of inter script similarities. In this work, we propose a new robust script identification system for Indian scripts that includes Indus documents and other scripts, namely, English, Kannada, Tamil, Telugu, Hindi and Gujarati which helps in selecting an appropriate OCR for recognition. The proposed system explores the spatial relationship between dominant points,namely, intersection points, end points and junction points of the connected components in the documents to extract the structure of the components. The degree of similarity between the scripts is studied by computing the variances of the proximity matrices of dominant points of the respective scripts. The method is evaluated on 700 scanned document images. Experimentalresults show that the proposed system outperforms the existing methods in terms of classification rate.*

***Keywords: Indus document, Dominant points, Proximity matrix, Variance, Indian scripts identification***

## 1.0INTRODUCTION

India is a multilingual country with documents of various languages and inscriptions. The earliest system of writing is found on seals known as Indus script used by the people of Indus valley from about 2500 B.C to 1500 B.C([14]; [13]). Indic scripts are evolved from the Brahmi script ([14]; [13]). A large numbers of Indus documents in seal form with corpora of texts in the form of symbols have been found in great cities Mohenjo-Daro and Harappa. Collection of such a large number of documents can be found in libraries and archaeology departments along with other language documents. The big question here is that; how to annotate or extract such documents as Indus scripts from a pool of mixed documents because Indus script is different from the other Indian scripts in terms of structure of text pattern, background, style of writing etc. This is valid because Indus scripts which were engraved on irregular surface sources such as rocks,may have any kind of background while other scripts such as English, Kannada, Telugu, Tamil, Hindi and Gujarati may have plain background and clear set of character patterns. For example, Fig. 1(a)-Fig. 1(g) show sample documents of Indus, English, Kannada, Tamil, Telugu, Hindi and Gujarati, respectively. It can be seen from Fig. 1(a) that text in Indus document has irregular structure components with complex background and is poor in quality.The presence of picture-like animal in Indus document makes problem more complex as compared to other scripts as shown in Fig. 1(b)-Fig. 1(g), which have plain background and regular structure of components. Therefore, it is an elusive goal for the researchers to develop a system for identifying Indus script along with other scripts. Besides, due to a huge collection of documents, it is difficult to identify and understand the scripts manually as it is time-consuming (i.e., several years to completion). In addition, lack of knowledge and experts of Indus scripts is another issue for annotation and automation.

Many methods ([1]; [4]; [5]; [15]) are developed in the past for recognizing Indian scripts, which generally focused on plain background images and used connected component analysis for recognition of scripts. The scope of the existing methods is limited to a particular script but not for multiple scripts because these methods exploited the knowledge of the scripts for script recognizing. In addition, there is no universal OCR available for recognizing multiple scripts. There are also methods which address the problem of poor quality text in low contrast images ([10]; [20]; [24]; [25]) for detection and recognition. However, these methods are developed for English text detection and recognition but not for other textssuch as Indus texts. It is noted from the methods ([10]; [20];




(a) Indus (b) English (c) Kannada (d) Tamil

(e) Telugu (f) Hindi (g) Gujarati

Fig.1: Samples documents images of different Indian scripts chosen from our database

[24]; [25]) that achieving high recognition rate for degraded and blurred text is still considered as an open issue in the field of video and camera-based image analysis. Therefore, we can conclude that these methods require individual scripts to achieve a high success rate for recognition. When the database contains documents of different scripts such as Indus documents, the recognition methods fail to recognize the text in the documents. Thus, script identification before recognition to select an appropriate OCR engine is essential, which helps in developing separate OCR for Indus documents.

We propose a robust script identification system for Indian historical scripts which includes Indus documents and other scripts. Hindi is the national language of India while Kannada, Tamil, Telugu and Gujarati are the official languages of Karnataka, Tamil Nadu, Andhra and Gujarat, respectively. We consider these scripts along with the Indus scripts for identification because these scripts are popular. Hence, there are higher chances of mixing them with Indus documents as reported bythe archeological department.

The paper is structured as follows. In section 2, we give a brief survey of related work. Section 3 is divided into three subsections to discuss the respective parts of the script identification system in detail, namely, dominant points extraction, feature extraction and template matching for identification.

## 2.0 RELATED LITERATURE

As discussed in the previous section, it is concluded that script identification is important for understanding documents before recognition. In this section, we provide a literature review on script identification methods. There are several methods proposed for script identification in scanned document images at the text line level and word level ([7]; [29]; [2]). For instance, an overview of script identification methodologies based on structure and visual appearance was designed by [7], where the proposed methods performed well for plain and high resolution camera-based images but not for documents like Indus scripts which usually have varying contrast, background, font shapes, font sizes, orientations and distortion.

Similarly, for the word level, there are methods ([21]; [3]; [12]) in which a good segmentation technique is required for segmenting words from text lines. Otherwise, segmentation error will affect the performance of the script identification. Meanwhile, there are methods which use shapes of characters for script identification (Shijian & Tan, 2008; [17]; Li & Tan, 2008). The main limitation of these methods is that the character shapes should be preserved to achieve a high script identification rate. However, it is observed that although methods for scanned and camera based images give high accuracy for degraded documents, the same methods may not be applicable for Indus script identification due to irregular structure of the character components and the presence of animal-like picture.

Furthermore, there are a few methods ([9]; [22]; [31]) for script identification in video to enhance the capability of OCR for recognizing multi-scripts. For example, the method proposed by [9] takes text lines as input and uses statistical and texture features with a k-nearest neighbor classifier to identify Latin and Ideographic text in images and videos. This method works well for high contrast English and Chinese text line images but not for other scripts. This is due to the fact that the texture features are extracted based on zonalization of text line images and the performance of the method is dependent on the number of the nearest neighbors used for classification. These features may not be good enough to identify low resolution scripts with complex background. New features, namely, smoothness and cursiveness based on text lines without a classifier are proposed by [22] for video script identification. The scope of the work is limited to two scripts presented in the document (i.e., English and Chinese or English and Tamil). To overcome this limitation, [31] proposed a method based on spatial-gradient-features to identify the script in a frame at the block level. Their proposal worked well for multi-script frames but not for a single frame containing multi-scripts because it expects one out of sixteen of the blocks to satisfy the condition for a particular script. In addition, using only fixed block size in this method may lead to loss of information. Another method of script identification and orientation detection for Indian text images were proposed by [8]. Since the features are derived from the connected component analysis of the characters, the performance of the method is good when the shape of the character is preserved. Recently, Sharma et al., 2013 proposed word-wise script identification in video at the word level. Their proposed method adopted the use of Zernike moments, Gabor features and gradient features with an SVM classifier. The method considered only Bangla, Hindi and English but not scripts like Indus, Kannada, Tamil and Gujarati, where a more powerful method is required.

In summary, from the above discussion on both scanned and video documents, it is observed that the main objective of the methods is to identify scripts such as printed text on different backgrounds but not handwritten text on complex backgrounds. Therefore, the methods may not work for Indus script identification due to the unpredictable nature of Indus text in Indus documents. As a result, we need a method that can work for both Indus documents and other scanned documents.

There area few methods for script identification in historical documents images ([23]; [18]; [11];[30]), which identify the script in Epigraphical document images. To remove noise in the Epigraphical document images, these proposed methods generally apply preprocessing steps (i.e., filtering) before extracting the features based on character shapes to identify the scripts. When a document contains more noise, the method fails to perform well because the method uses simple filtering methods. Moreover, the extracted features may overlap with the features of picture-like animals in the Indus document. Therefore, these methods do not have the ability to identify scripts in Indus document. Recently, Indus document classification from other script documents has been proposed by [14]; [13], where these methods utilized straightness and cursiveness of components to classify Indus and English documents. Since the objective of the method is to classify Indus and English, the same features lose discriminative power because straightness and cursiveness also exist in all other Indian scripts. Therefore, these features may not be good enough to classify scripts like Kannada, Telugu, Tamil, Hindi and Gujarati.

This motivates us to propose a new method based on spatial study of text components to classify Indus scripts and other scripts in this work. The proposed method works based on the fact that the structure of each component differs from one script to another script. To extract the structures of all the components, we explore the spatial relationship between dominant points, namely, end points, junction points and intersection points. The degree of similarity is estimated using the variances of distance matrices of the respective dominant points to classify the scripts.

## 3.0THE PROPOSED METHOD

We propose a method known as Variance Dominant Pixel method (VDM) for identifying seven scripts, namely, Indus, English, Kannada, Telugu, Tamil, Hindi and Gujarati based on the spatial relationship between the dominant pixels of text components in the scanned document images. Inspired by the work presented by [28] for video script identification, we propose features extraction based on the proximity between dominant pixels, namely intersection points, junction points and end points of text components. This is owing to the fact that these features extract the structure of the texts pattern, which subsequently give clues to differentiate the above scripts. It is known that the structures of the text components differ from one script to another script. For example, cursive structures of the components can be seen in Indus documents due to the variation in writing style. The other scripts, namely, Kannada, Tamil, Telugu, Hindi and Gujaratihave more cursive components as compared to the English script.Hence, we expect more end points, junction points and intersection points. This

is the main fundamental for proposing features to identify the seven scripts. To extract such observations, VDM converts an input image to binary image and then applies a thinning algorithm to reduce each stroke width to a single pixel width. For each component in a text line, the method extracts dominant pixels and finds proximity between them based on Euclidean distance. The variance of proximity matrix of each pixel is computed. Finally, the average of variances of the components of the text line is considered as a feature for discriminating the scripts.

The method extracts three features with respect to end points, junction points and intersection points. The feature vector consisting of the three features is then matched with a set of pre-defined templates based on the training samples for the seven scripts to be identified. A salient point to note here is that the three features are all that we need to identify the seven scripts rather than the much large number of features used in [28]. This is the main contribution of the work. The block diagram of the proposed method as shown in Fig. 2 shows the distinctive steps of our proposed method. The proposed method is divided into three sub-sections. Section 3.1 describes how to obtain dominant pixels. Section 3.2 shows the features applied on the dominant pixels. Section 3.3 describes the template creation to classify scripts.

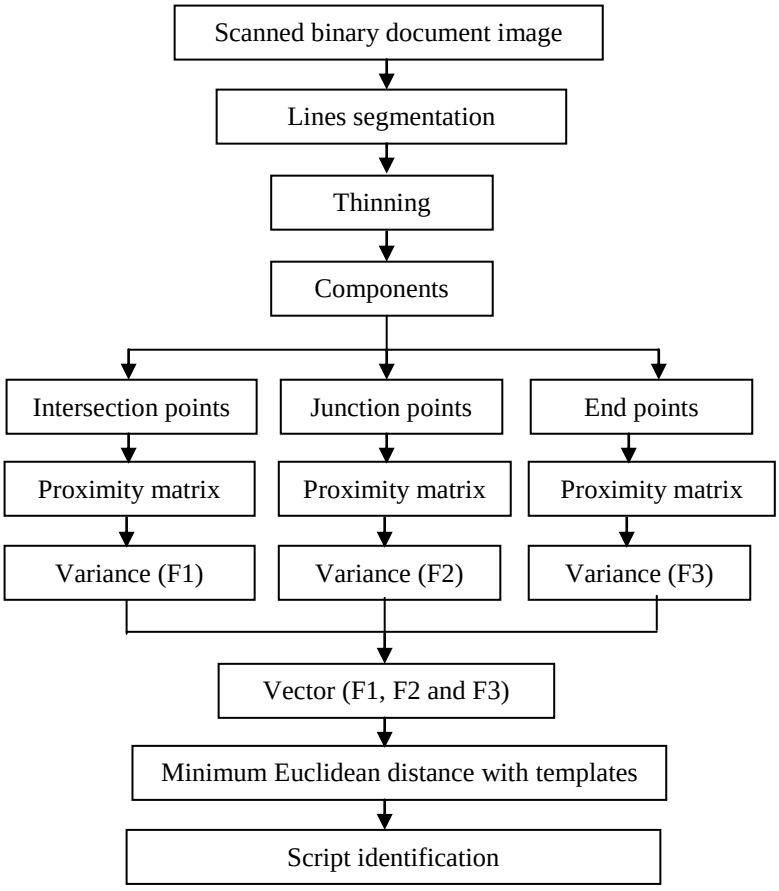



Fig. 2: Logical flow of the proposed method

### 3.1. Extraction of Dominant Pixels

From the scanned document image, our proposed method obtains a binary image and then segments the text lines using a region growing method. Our proposal segments text lines based on the nearest neighbour criterion as proposed by [14]. The method uses region growing for extracting features and segmenting text lines. It selects text lines which have a larger number of character components for feature extraction. To reduce the stroke width

of the components, the method adopts a thinning process used in the work proposed by [13]to find the exact intersection points, junction points and end points for each character in the text line. In addition, it also helps in reducing the number of computations. The method finds intersection points, junction points and end points based on the neighbouring pixels while traversing along the contour of the components. The conditions for defining intersection points, junction points and end points are dependent on the number of neighbouring pixels to be either 4, 3 or 1, respectively. As a result, we get three different kinds of dominant pixels for a component as shown in Fig. 3, where one can see pixels representing intersection points, junction points and end points.

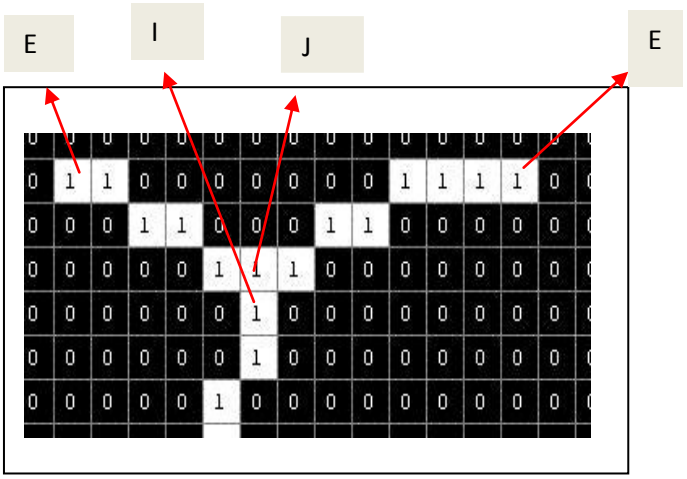
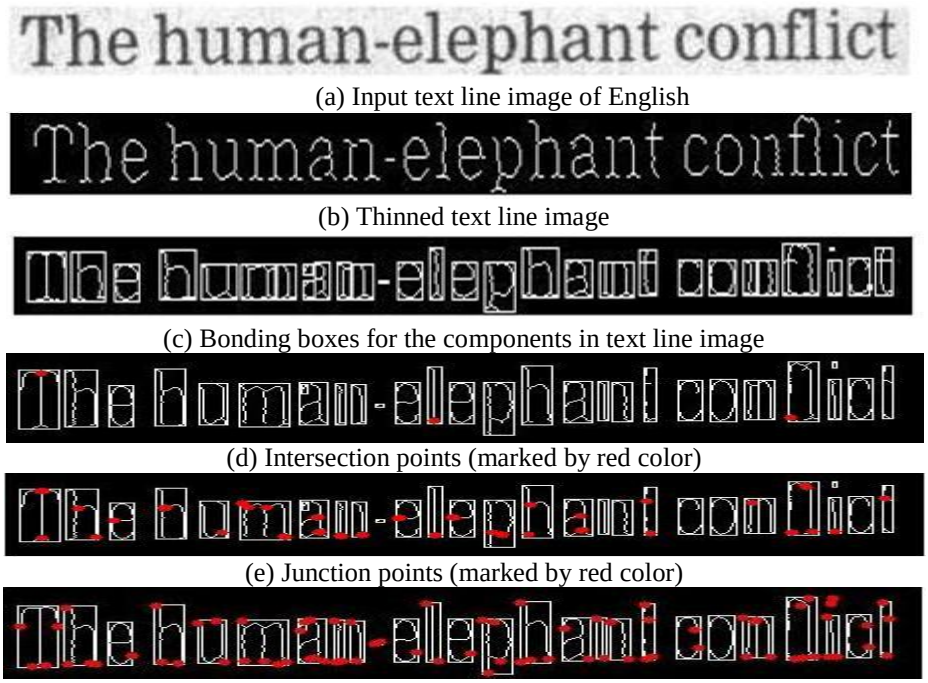



Fig. 3: Dominant pixels selection for each component: I- Intersection points, J-Junction points, E-End points.

The process for extracting dominant pixels from a text line is illustrated in Fig. 4, where (a) shows the English input text line, (b) shows the results of thinning, (c) shows bounding boxes for each of the components in the text line, (d) shows the result of intersection points, marked in red, (e) shows the result of junction points, marked in red and (f) shows the result of end points, marked in red. It is noticed from Fig. 4(d)-(f) that there are more end points as compared to intersection points and junction points in English text because generally connected components of English text have less cursiveness as compared to the other scripts.We can see a lot of intersection points, junction points as well as end points in other scripts. This makes a significant difference in discriminating English from the other scripts. Fig. 5 shows clearly each point selection for the text components.



(a) Input text line image of English

(b) Thinned text line image

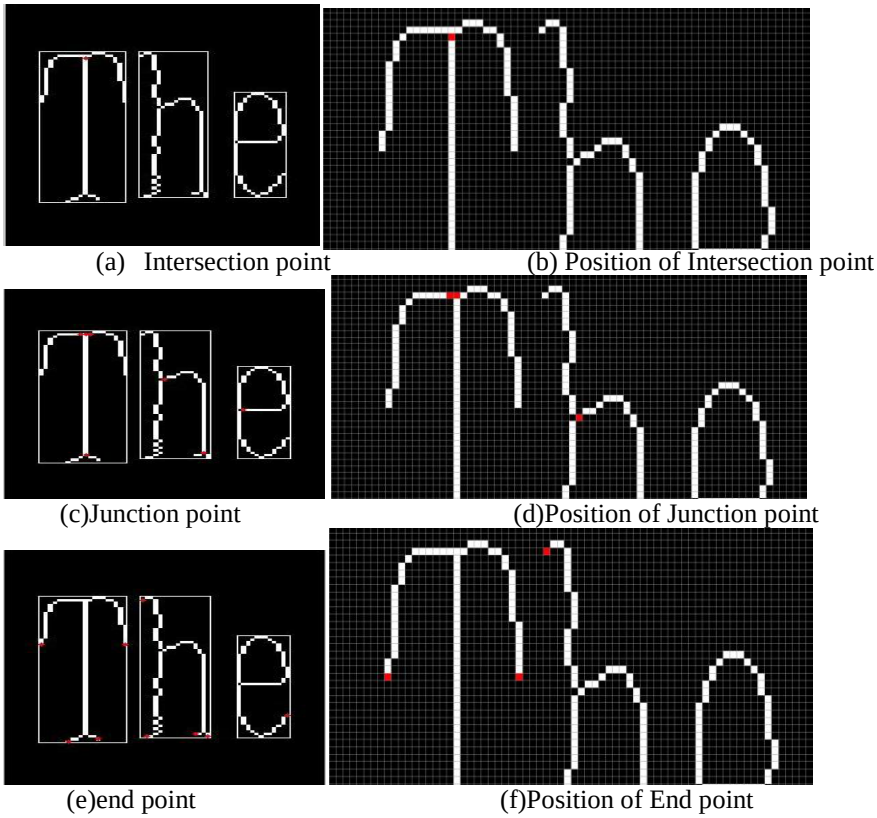
(c) Bonding boxes for the components in text line image

(d) Intersection points (marked by red color)

(e) Junction points (marked by red color)

(f) End points (marked by red color)

Fig. 4: Intermediate steps for the dominant pixel extraction



(a) Intersection point (b) Position of Intersection point

(c)Junction point (d)Position of Junction point

(e)end point (f)Position of End point

Fig. 5: Position of dominant pixel(marked in red)

### 3.2. Features Extraction from Dominant Pixels

In order to study the proximity between dominant pixels which gives spatial relationship between the pixel and the structure of the components, our proposed method estimates the distance from the first dominant pixel to all the remaining dominant pixels, say distance-1 and then from the second dominant pixel to all the remaining dominant pixels including the first dominant pixel, say distance-2 and so on.

As a result, we get a proximity matrix for each set of dominant pixels of each component in the text line. The method computes the variances of the proximity matrices corresponding to the three sets of dominant pixels of each component to estimate the degree of similarity of the dominant pixels distribution in the components. Finally, the method computes the average of variances of each component in the text line as a feature for discrimination. More specifically, let$(x1, y1)$ and $(x2, y2)$denote the coordinates of two dominant pixels. Mathematically, the proximity between the points is obtained by:

$$\text{Euclid\_dist}(x1, x2) = \text{SQRT}((x2 - x1)^2 + (y2 - y1)^2) \quad (1)$$

Proximity matrix (distance matrix) is obtained for each of the connected component as follows.

$$\text{Dist\_matrix}(i, j) = \sum_{\text{i=1,j=1}}^{n} \text{Euclid\_dist}(i, j) \quad (2)$$

Let distancematrices obtained from intersection points, junction points and end points be represented as *IDM*(*i*,*j*), *JDM*(*i*,*j*),*EDM*(*i*,*j*), respectively. Variances of different distance matrices for each component are computed as follows:

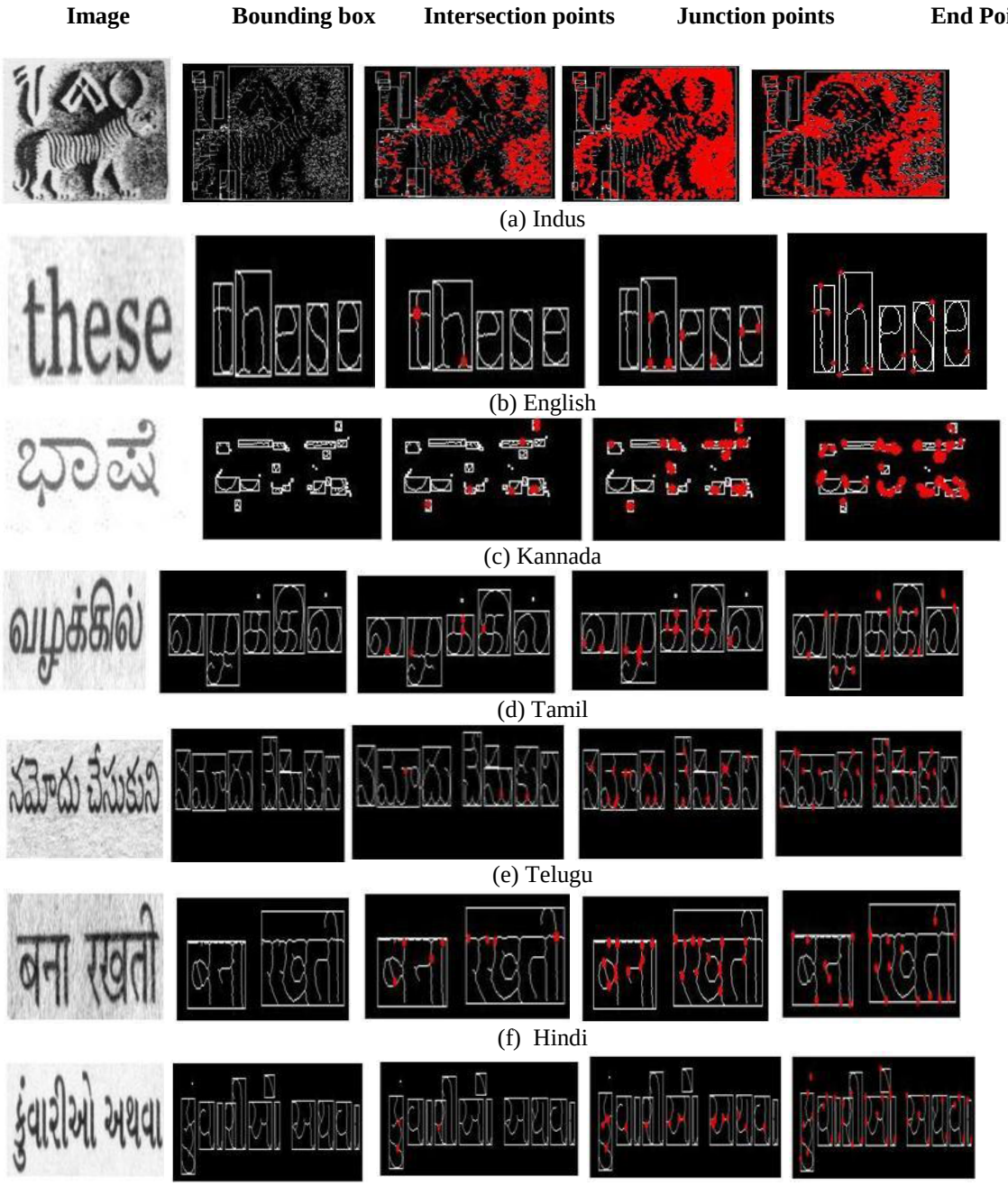
$$\begin{aligned} \text{CC}_{\text{ii}} &= \text{VAR}\big(\text{IDM(i, j)}\big) \\ \text{CC}_{\text{ij}} &= \text{VAR}\big(\text{JDM(i, j)}\big) \quad (3) \\ \text{CC}_{\text{ie}} &= \text{VAR}\big(\text{EDM(i, j)}\big), \end{aligned}$$

where$CC_{ii}$, $CC_{ij}$, $CC_{ie}$ correspond to variances for intersection points, junction points and end points respectively. The method again computes the mean for all the variances of the components in a text line as given below. Let F1, F2, F3 denote the features for intersection points, junction points and end points, respectively. The feature values for the whole line containing all the componentsare calculated as follows,

$$\begin{aligned} \text{F1} &= \frac{1}{n}\sum_{\text{k=0}}^{\text{n}} \text{CC}_{\text{ii}} \\ \text{F2} &= \frac{1}{n}\sum_{\text{k=0}}^{\text{n}} \text{CC}_{\text{ij}} \quad (4) \\ \text{F3} &= \frac{1}{n}\sum_{\text{k=0}}^{\text{n}} \text{CC}_{\text{ie}} \end{aligned}$$

where*n* denotes the number of components in a text line.

The three features (F1, F2 and F3) for the seven scripts are shown in Fig. 6 where (a)-(g) show the three dominant pixels extraction and representation for the seven scripts, respectively. It is observed from Fig. 6 that the structures of dominant pixel representations for the seven scripts exhibit clear distinction among the scripts. For example, the dominant pixels distribution of the Indus scriptas shown inFig. 6(a) shows that the pixels are much closer to each other as compared to the other scripts due to the cursiveness and complex background of the Indus text. In this way, the method extracts three distinct features for classification of the seven scripts.




(a) Indus

(b) English

(c) Kannada

(d) Tamil

(e) Telugu

(f) Hindi

(g) Gujarati

Fig.6: Three features with respect to three dominant pixels for the seven scripts

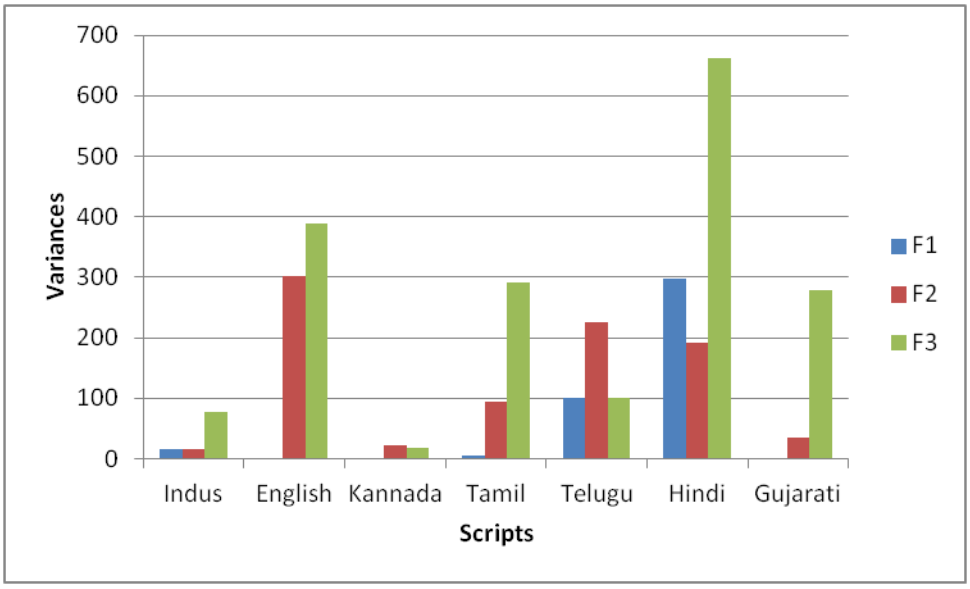



Fig. 7: Three features representation for the seven scripts of Fig.6

The discriminative power of the three features for the seven scripts shown in Fig. 6 can be observed in Fig. 7. It is obvious thatthe range of variances differs for different scripts. It is observed that some of the feature values are similar for different scripts. Irregular surface structure in Indus text results in many components and in turn their values are minimum compared to other scripts. However due to the animal-like structure and cursive text,there are more dominant feature points compared to text images of other scripts. Scripts like Kannada contains subcomponents for each text component as shown in Fig. 6(c), resulting in minimal end components. Hence these three features are effective for classification of the seven scripts.

### 3.3. Template Construction for Script Classification

We choose the training samples according to a five-fold cross validation criterion from each of the seven scripts database for constructing the templates. The feature vector of an unknown script is then compared with the template created from the training samples. We consider each script database as one class and hence this is a seven-class classification problem.Motivated by the work presented by [17] and [31] for script identification, we compute the averages of feature vectors from the training samples chosen according to the five-fold cross validation criterion for the respective seven classes. Examples in Fig. 8 shows50 samples chosen randomly from the respective script databases. It can be seen from Fig. 8 that all the templates have different distributions. This is the advantage of our proposed method for script identification. Let the variance features (F1, F2 and F3) obtained for intersection points, junction points and end points be VAR_I, VAR_J and VAR_E, respectively. The averages of these variance feature vectors are computed as defined in equation (5), resulting in seven templates, respectively as AI,AJ and AE.

$$\mathrm{AI} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=0}^{\mathrm{m}-1} \mathrm{VAR\_I(i)} \quad , m = 50$$

$$\mathrm{AJ} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=0}^{\mathrm{m}-1} \mathrm{VAR\_J(i)}, m = 50 \qquad (5)$$

$$\mathrm{AE} = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=0}^{\mathrm{m}-1} \mathrm{VAR\_E(i)} \quad , m = 50$$

To classify a script,VDM extracts features (i.e., F1, F2 and F3) for each connected component as shown in equation (4). Next, it finds the minimum Euclidean distance as shown in equation (6) between features of the text line and the corresponding templates. Let $T_i$= {$T1_i$...$T7_i$} denotes the set oftemplates generated as in Fig. 8 for all the seven scripts for feature $F_i$.

$$E\_Dist = SQRT(ABS(T_i - F_i)^2) \quad (6)$$

The template that gives the minimum E_Dist determines the script type. Fig.8 shows that the templates have distinct distribution for each script. Therefore, we can infer that the extracted features are sufficient for classification of scripts.

## 4.0 EXPERIMENTAL RESULTS

We create our own database for experimentation and evaluation because there is no standard or databases available asbenchmark. Our database consists of images from different sources, such as newspaper, magazine, books etc. Particularly, Indus documents are collected from an Archeological survey done in Mysore Karnataka, India. We obtain one hundred images for each script giving rise to a total of seven hundred documents for the purpose of experimentation. We use classification rate for evaluating the method as detailed in [28]. The classification rate is computed based on the diagonal elements in the confusion matrix. The confusion matrices are obtained based on the five-fold cross validation experimentations. The final classification rate is computed by taking the average of all the five-fold confusion matrices. To assess the effectiveness of the proposed method, we compare the proposed method with several existing methods in terms of classification rate. The existing methods include [19] which used projection profile-based features for Indian script identification; [6] which used connected component analysis for south Indian script identification and [11] which used texture features for script identification in handwriting document images. The reason to choose these methods is that they were developed for Indian script identification. In addition, these methods used connected component analysis-based features and texture features for classification of scripts. On the other hand, our proposed method used spatial information and structure of the text pattern for classifying scripts. On top of this, the method considers Gujarati and Hindi for classification as these two documents appear to be the same and share the same properties. Furthermore, to test the robustness of thinning which is used in VDM for classification, we also evaluate the confusion matrix generated for Sobel and Canny edge images as in the proposed method but without thinning.

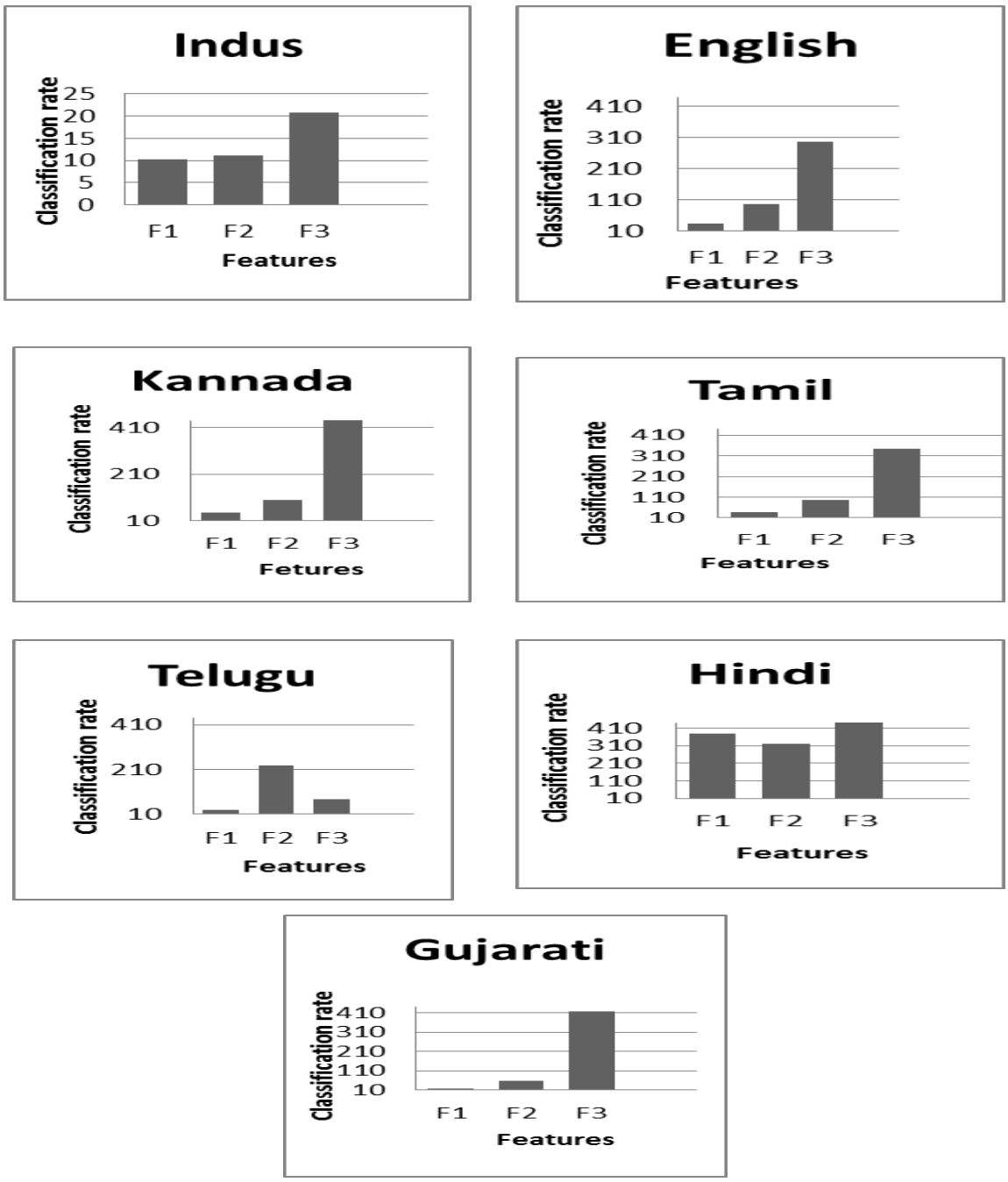



Fig. 8: Templates generated for different scripts

To understand the contribution of each features among the three features, we calculate the classification rate for each feature individually.The classification rate for eachfeature is shown in Fig. 9.It is noticed that F3 contributes more as compared to F2 and F1. In this experiment, we use the templates shown in Fig. 8 which were created based on the average of 50 training samples chosen randomly from each of the script databases. The result shows that the collective contribution gives good classification rate.

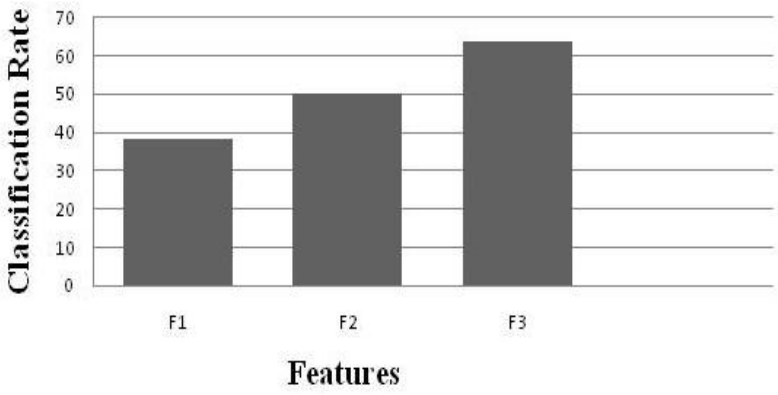



Fig.9: Individual feature analysis

### 4.1. Performance Evaluation

The quantitative results of the proposed method are reported in Table 1 without five-fold cross validation criterion, where the method is tested on all documents. In this experiment, we use the templates created by taking an average of 50 training samples chosen randomly from databases. Then, all documents of the respective classes are compared with the respective templates for classification. Therefore, this gives good results for all the scripts because of more number of samples and there is no separation of testing samples. Overall, the method gives 96.1 % classification rate. We use MATLAB 2011to evaluate the proposed algorithm. The average processing time for each script is 50.4, 0.32, 0.61, 0.8, 0.3, 0.4, 0.5 millisecond for Indus, English, Kannada, Tamil, Telugu, Hindi and Gujarati, respectively. Average running time is measured in Intel® Core i3 processor M330@2.13GHZ, 914 MHz, 2.92 GB RAM.

Table1. Confusion matrix for the proposed method in %

| Script type | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **98** | 0 | 1 | 0 | 0 | 1 | 0 |
| English | 0 | **97** | 0 | 2 | 1 | 0 | 0 |
| Kannada | 0 | 0 | **97** | 0 | 0 | 0 | 3 |
| Tamil | 0 | 3 | 0 | **93** | 1 | 0 | 3 |
| Telugu | 0 | 1 | 0 | 0 | **98** | 0 | 1 |
| Hindi | 1 | 0 | 4 | 0 | 0 | **95** | 0 |
| Gujarati | 0 | 0 | 2 | 2 | 1 | 0 | **95** |

### 4.2. Fivefold Cross Validation for Evaluation

In order to evaluate the strength of the method, we conduct experiments based on standard five-fold cross validation in terms of classification rate. We repeat this experiment by dividing the whole dataset into five subsets. Each time one out ofthe five subsets is considered as the training set to create the template while the remaining four subsets are considered as test data. The results for each fold experiment are reported in Table 2-Table 6. Table 7 gives the average of the diagonal values from Table 2-Table 6. The average of diagonal elements in Table 7 gives 95 % classification rate. When we compare Table 1 to Table 7, classification rate of the Table 1 is 96.1% the difference is very negligible. One can observe from Table 2-Table 6 that different ways of experimentation does not affect the results much.The results in Table 2-Table 6 aresimilar and there is no drastic change in the results. It is evident that the extracted features are stable and it can be extended to a larger number of scripts.

Table 2. Confusion matrix generated during first fold

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **97** | 0 | 1 | 0 | 1 | 1 | 0 |
| English | 0 | **96** | 1 | 2 | 1 | 0 | 0 |
| Kannada | 0 | 0 | **97** | 0 | 0 | 0 | 3 |
| Tamil | 0 | 5 | 0 | **91** | 1 | 0 | 3 |
| Telugu | 0 | 2 | 0 | 0 | **98** | 0 | 0 |
| Hindi | 1 | 0 | 4 | 0 | 0 | **95** | 0 |
| Gujarati | 0 | 0 | 2 | 3 | 1 | 0 | **94** |

Table 3. Confusion matrix generated during second fold

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **96** | 1 | 1 | 0 | 1 | 1 | 0 |
| English | 0 | **97** | 1 | 2 | 0 | 0 | 0 |
| Kannada | 0 | 0 | **96** | 0 | 0 | 0 | 4 |
| Tamil | 0 | 4 | 0 | **90** | 1 | 0 | 5 |
| Telugu | 0 | 8 | 0 | 2 | **90** | 0 | 0 |
| Hindi | 1 | 0 | 5 | 0 | 0 | **94** | 0 |
| Gujarati | 0 | 0 | 2 | 3 | 0 | 0 | **95** |

Table 4. Confusion matrix generated during third fold

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **97** | 0 | 1 | 0 | 1 | 1 | 0 |
| English | 0 | **97** | 0 | 2 | 1 | 0 | 0 |
| Kannada | 0 | 0 | **97** | 0 | 0 | 0 | 3 |
| Tamil | 0 | 5 | 0 | **91** | 1 | 0 | 3 |
| Telugu | 0 | 2 | 0 | 2 | **96** | 0 | 0 |
| Hindi | 1 | 0 | 2 | 0 | 0 | **97** | 0 |
| Gujarati | 0 | 0 | 2 | 2 | 0 | 0 | **96** |

Table 5. Confusion matrix generated during fourth fold

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **98** | 0 | 1 | 0 | 0 | 1 | 0 |
| English | 0 | **95** | 1 | 3 | 1 | 0 | 0 |
| Kannada | 0 | 0 | **97** | 0 | 0 | 0 | 3 |
| Tamil | 0 | 5 | 0 | **91** | 1 | 0 | 3 |
| Telugu | 0 | 4 | 0 | 2 | **94** | 0 | 0 |
| Hindi | 1 | 0 | 6 | 0 | 0 | **93** | 0 |
| Gujarati | 0 | 1 | 1 | 2 | 0 | 0 | **96** |

Table 6. Confusion matrix generated during fifth fold

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **97** | 0 | 1 | 0 | 1 | 1 | 0 |
| English | 0 | **97** | 0 | 2 | 1 | 0 | 0 |
| Kannada | 0 | 0 | **93** | 0 | 0 | 0 | 7 |
| Tamil | 0 | 3 | 0 | **93** | 1 | 0 | 3 |
| Telugu | 0 | 2 | 0 | 0 | **98** | 0 | 0 |
| Hindi | 1 | 0 | 6 | 0 | 0 | **93** | 0 |
| Gujarati | 0 | 1 | 3 | 1 | 0 | 0 | **95** |

Table 7. Average classification rates

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **97** | 0.2 | 1 | 0 | 0.8 | 1 | 0 |
| English | 0 | **96.4** | 0.6 | 2.2 | 0.8 | 0 | 0 |
| Kannada | 0 | 0 | **96** | 0 | 0 | 0 | 4 |
| Tamil | 0 | 4.4 | 0 | **91.2** | 1 | 0 | 3.4 |
| Telugu | 0 | 3.6 | 0 | 1.2 | **95.2** | 0 | 0 |
| Hindi | 1 | 0 | 4.6 | 0 | 0 | **94.4** | 0 |
| Gujarati | 0 | 0.4 | 2 | 2.2 | 0.2 | 0 | **95.2** |

### 4.3. Comparative Study

The confusion matrices for Sobel edge and Canny edge images are shown in Table 8 and Table 9, respectively. The classification rates of the proposed method and existing methods are reported in Table 10. Table 10 shows that the method which uses Canny operation outperforms other existing methods and the method which uses Sobel operation. This is because Sobel operator does not produce edges appropriately for low resolution images with complex background and hence loses the shape structures of the characters.Meanwhile, Canny operator preservesthe shapes of the characters. However, sometimes Canny operator gives more spurious edges due to complex background. This leads to low classification rate as compared to the proposed method. Scheme proposed by [19]extracted top-bottom curves, pipe density. The scope is limited to south Indian and English scripts and hence it cannot handle all the scripts that we consider here.Therefore, their proposal gives poorer results as compared to other methods. Algorithm developed by [6]performs well on structure of components. Structure is not retained as exactly as it appears in the scanned documents. Hence their scheme provides 47 % classification rate. The proposal of [11]works on various texture features such as mean, variance, entropy, smoothness etc. These features may not be sufficient for classification of all the seven scripts. Their proposal achieved an overall 44 % of classification rate (i.e., as shown in Table 10). Thus, we can conclude that our proposed method is more effective and robust as compared to the existing methods.

Table 8. Confusion matrix generated for Sobel edges

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **83** | 0 | 0 | 1 | 0 | 1 | 0 |
| English | 0 | **9** | 1 | 79 | 2 | 0 | 9 |
| Kannada | 0 | 0 | **60** | 0 | 0 | 40 | 0 |
| Tamil | 0 | 6 | 3 | **48** | 3 | 0 | 40 |
| Telugu | 4 | 23 | 0 | 26 | **45** | 0 | 2 |
| Hindi | 4 | 3 | 17 | 1 | 2 | **67** | 6 |
| Gujarati | 0 | 0 | 78 | 0 | 0 | 20 | **2** |

Table 9. Confusion matrix generated for Canny edges

| | Indus | English | Kannada | Tamil | Telugu | Hindi | Gujarati |
|---|---|---|---|---|---|---|---|
| Indus | **98** | 1 | 0 | 0 | 0 | 1 | 0 |
| English | 0 | **7** | 0 | 1 | 91 | 0 | 1 |
| Kannada | 0 | 0 | **48** | 6 | 0 | 0 | 46 |
| Tamil | 0 | 41 | 0 | **34** | 23 | 0 | 2 |
| Telugu | 6 | 2 | 0 | 2 | **92** | 0 | 0 |
| Hindi | 5 | 7 | 27 | 6 | 6 | **32** | 17 |
| Gujarati | 0 | 0 | 4 | 14 | 1 | 0 | **81** |

Table10. Classification rate of the proposed method and existing methods

| **Methods** | **Classification rate in %** |
|---|---|
| [19] | 7 |
| [6] | 47 |
| [11] | 44 |
| Proposed method-Sobel edge (Table 8) | 44.8 |
| Proposed method-Canny edge (Table 9) | 56 |
| **Proposed Method-Thinning** | **96.1** |

For some cases as shown in Fig. 10, our proposed method misclassifies due to overlapping features for the document images where blurring exists thus degrading the image quality. Since classifying the seven scripts is a complex problem, the proposed method gives poor results for some cases. For example, Indus script may be misclassified as Kannada as the text pattern looks similar.Similarproblem occurs when Indus is misclassified as Hindi. Therefore, there is scope for improvement in future. We will investigate further methods to strengthen our proposal in order to improve the classification rate.


Indus as Kannada Indus as Hindi English as Telugu English as Hindi

Kannada as Tamil Kannada as Gujarati Hindi as Kannada Hindi as Gujarati

Tamil as English Tamil as Gujarati Telugu as English Telugu as Gujarati

Hindi as English Hindi as Kannada Gujarati as Kannada Gujarati

Fig.10: Example for misclassification

## 5.0 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new script identification system for classification of Indus scripts, English texts and five other Indian scripts, namely, Kannada, Telugu, Tamil, Hindi and Gujarati. The proposed method is designed based on the proximity between dominant pixels, namely intersection points, end points and junction points of the connected components found in the documents. The advantage of these dominant points is the distance between these points helps us to understand the structure of the components. The degree of similarity is estimated based on the variances of the proximity matrices. Experimental results show that the thinning process of the proposed method plays an important role in script classification.Experimental results show that the proposed method outperforms the existing methods in terms of classification rate. In summary, the following are the contributions and limitations of the proposed system.

Contributions:

1. Explore the distances between pairs of dominant points to extract the structures of the connected components contributing to the identification of different scripts especially for Kannada-Telugu-Tamil and Hindi-Gujarati as these scripts look similar in the structures of their components.

2. Propose a set of simple yet effective features, based on the variances of the proximity matrices indifferent scripts for studying the degree of similarity between unknown script and the templates.

3. Solve the complex script identification problem without using an expensive classifier and a large number of training samples.To evaluate the robustness of the proposed system for identifying scripts in historical documents like Indus scripts where there area lot of distortions due to non-rigid surface and usual style of handwriting

Limitations:

1. As shown in Fig. 10, the proposed method misclassifies the scripts when the input document is blurred and too noisy. Therefore, our next target is to enhance the proposed system for identifying blurred and noisy documents.

2. In this work, the scope is limited to seven scripts. Our future work will focus on improving the proposed system by including international scripts and adding a new set of features.

3. The current proposed system is tested on off-line data. Our next task is to test the proposed system on online data for real time applications.

4. The current work requires an entire text line for script identification. Sometimes, a single text line may contain words of different scripts. Therefore, there is a scope for extending this system for identifying scripts at the word level.

5. Our next target is to propose a new method for recognizing Indus text to develop an automatic system for recognizing and understanding historical documents.

**Acknowledgement**

The work is partly supported by the University of Malaya HIR under Grant No. M.C/625/1/HIR/210.

## REFERENCES

[1] T. V. Ashwin and P. S. Sastry, “A font and size-independent OCR system for printed Kannada documents using support vector machines”. *Sadhana*, Vol. 27, No. 1, February 2002, pp. 35-58.

[2] A.Busch, W. W. Boles and S. Sridharan , "Texture for Script Identification". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No.11, IEEE Computer Society, 2005, pp. 1720-1732.

[3] S. Chanda, S.K.Pal, K. Franke and U. Pal, "Two-stage Approach for Word-wise Script Identification". *International Conference on Document Analysis and Recognition*, IEEE, July 2009, pp. 926-930.

[4] B. B. Chaudhuri and U.Pal, " An OCR System to Read Two Indian Languages Scripts: Bangla and Devanagari". *International Conference on Document Analysis and Recognition*, Aug 1997, pp.1011-1015.

[5] B. B. Chaudhuri and U. Pal, "A Complete Printed Bangla OCR System". *Pattern Recognition*,Vol. 31, No. 5, March 1998, pp.531-539.

[6] M. S. Das, D. S. Rani, C. R. K. Reddy and A. Govardhan , "Script identification from Multilingual Telugu, Hindi and English Text Documents". *International Journal of Wisdom Based Computing* , Vol. 1, No. 3, 2011

[7] D. Ghosh , T. Dube and A.P. Shivaprasad , " Script Recognition-Review". *IEEE Transactions on Pattern Analysis & Machine Intelligence* ,Vol.32, December 2010 , pp.2142-2161.

[8] S. Ghosh and B. B.Chaudhuri , " Composite Script Identification and Orientation Detection for Indian Text Images". *International Conference on Document Analysis and Recognition, IEEE*, September 2011, pp.294-298.

[9]J.Gllavata and B.Freisleben, "Script Recognition in Images with Complex Backgrounds". *Signal Processing and Information Technology*, IEEE, 2005, pp. 589-594.

[10] M.Grafmuller and J. Beyerer, "Performance improvement of character recognition in industrial applications using prior knowledge for more reliable segmentation". *Expert Systems with Applications*, Vol. 40, No. 17, 2013, PP. 6955 – 6963.

[11] M.Hangargea , K. C. Santoshb, S. Doddamania and R.Pardeshia, "Statistical Texture Features based Handwritten and Printed Text Classification in South Indian Documents". *International Conference on Emerging Computation & Information Technologies*,Elsevier, 2013, pp. 215-221

[12] M. A. Shayegan, S. Aghabozorgi, and R. G. Raj, "A Novel Two-Stage Spectrum-Based Approach for Dimensionality Reduction: A Case Study on the Recognition of Handwritten Numerals," Journal of Applied Mathematics, vol. 2014, Article ID 654787, 14 pages, 2014. doi:10.1155/2014/654787.

[13] A. S.Kavitha, P. Shivakumara and G.H. Kumar " An Integrated Method for Classification of Indus and English Document Images". *International Conference on Emerging Research in Electronics, Computer Science and Technology*, Springer, Vol. 248, 2014, pp.343-355.

[14] A. S. Kavitha , P. Shivakumara and G.H. Kumar, "Skewness and Nearest Neighbour Based Approach for Historical Document Classification". *International Conference on Communication Systems and Network Technologies*, IEEE, April 2013, pp. 602-606.

[15] P.Krishnan, N. Sankaran , A. K. Singh and C. V. Jawahar , "Towards a robust OCR system for Indic scripts". *Document Analysis Systems*, IEEE ,April 2014,pp.141-145.

[16] L. Li, and C. L. Tan, "Script Identification of Camera-based Images". *International Conference onPattern Recognition*, IEEE , December 2008, pp. 1 – 4.

[17] S.Lu, L. Li and and C. L. Tan, "Identification of scripts and orientations of degraded document images". *Pattern Analysis and Applications*, Springer, Vol. 13 , No. 4, November 2010, pp.469-475.

[18] K. S. Murthy, G. H. Kumar, P.Shivakumara and P. R. Ranganath, "Nearest Neighbour Clustering approach for line and character segmentation in epigraphical scripts". *International Conference on Cognitive Systems*,2004

[19] M. C. Padma and P. A. Vijaya " Script identification from trilingual documents usingprofile based features". *International Journal of Computer Science and Applications*, Vol. 7, No. 4, 2010, pp.16 – 33.

[20] J. G. Park and K. J. Kim, " Design of a visual perception model with edge-adaptive Gabor filter and support vector machine for traffic sign detection". *Expert Systems with Applications*, Elsevier , Vol. 40, No.9, July 2013, pp. 3679 – 3687.

[21] P. B.Pati, and A. G. Ramakrishnan , "Word level multi-script identification". *Pattern Recognition Letters*, Elsevier, Vol. 29, No. 9, July 2008, pp.1218-1229.

[22] T. Q. Phan, P.Shivakumara, Z. Ding, S. Lu and C. L. Tan, "Video Script Identification based on Text Lines". *International Conference on Document Analysis and Recognition*, September 2011, pp.1240-1244.

[23] S.Rajkumar and S. Bharathi, "Ancient Tamil Script Recognition from Stone Inscriptions Using Slant Removal Method".*IEEE International Conference on e-Business Engineering*, 2012

[24] A.Risnumawan, P. Shivakumara, C. S. Chan and C. L.Tan , "A Robust Arbitrary Text Detection System for Natural Scene Images". *Expert Systems with Applications*, Elsevier,Vol. 41, No. 18 ,pp. 8027-8048.

[25] S.Roy, P.Shivakumara, P. P. Roy, U.Pal,C.L.Tan, T.Lu, "Bayesian Classifier for Multi-Oriented Video Text Recognition System". *Expert Systems with Applications* , Elsevier, March 2015

[26] N.Sharma, S.Chanda, U. Pal and M.Blumestiein,"Word-Wise Script Identification from Video Frames". *International Conference on Document Analysis and Recognition*,IEEE,2013 , pp.867-871.

[27] L. Shijian and C.L. Tan, " Script and Language Identification in Noisy and Degraded Document Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*,Vol.30, No. 1, January 2008, pp. 14-24.

[28]P. Shivakumara, Z. Yuan,D. Zhao,T. Lu and C. L.Tan, " New Gradient-Spatial-Structural-Features for Video Script Identification". *Computer Vision and Image Understanding*, Elsevier, Vol. 130,January 2015, PP.35-53.

[29] T.N.Tan (1998). Rotation Invariant Texture Features and Their Use in Automatic Script Identification. *IEEE Transactions on* Pattern Analysis and Machine Intelligence.Vol. 20, No.7, *July 1998* pp.751-756.

[30] Moohebat, M., Raj, R.G. , Kareem, S.B.A., Thorleuchter, D., "Identifying ISI-indexed articles by their lexical usage: A text analysis approach", Journal of the Association for Information Science and Technology, Vol. 66, No. 3, pp. 501–511. doi: 10.1002/asi.23194.

[31] D.Zhao,P. Shivakumara, S.Lu and C.L.Tan, "New Spatial-Gradient-Features for Video Script Identification". Document Analysis Systems,IEEE,March 2012, PP. 38-42.